

# Analyzing Genomic Data with NOJAH

## TAB A) GENOME WIDE ANALYSIS



A Genome-Wide Heatmap can be very dense. Given the limitation with the computational power required to construct a genome wide heatmap, NOJAH showcases a [Genome-Wide Dendrogram](#).

**Genome-Wide Heatmap Analysis workflow is divided into four main subparts:**

1. [Identify the Most Variable Features](#)
2. [Construct a HeatMap for the Most Variable Features](#)
3. [Identify Number of Clusters and Assess Cluster Stability](#)
4. [Identify Core Samples](#)

Heatmap is [updated](#) based on the Consensus Core Samples.

However each of these components are not dependent on each other and can be used independently.

The 'Input GW Data' form has a title 'Input GW Data' and a subtitle 'Select an example dataset or upload your own with 'Load my own GW data.''. It features a dropdown menu with 'Example TCGA BRCA Exp Data' selected. Below the dropdown is a button labeled 'Download GW TCGA BRCA Expression DataSet'.

### Step 1:

Select the example dataset or upload your own. Two example datasets are available. Genome-Wide TCGA-BRCA Expression datasets and CoMMpass RNASeq Expression dataset are available.

To view example data and format, use download button to view CSV file.

The 'Data Subsetting' form has a title 'Data Subsetting' and a subtitle 'Subset GW data by:'. It includes three checkboxes: 'Variance' (unchecked), 'Median Absolute Deviation' (unchecked), and 'Inter Quartile Range' (checked). Below these is a 'Percentile' section with two radio buttons: 'Percentile Slider' (selected) and 'Manually Enter Percentile' (unselected). A 'Percentile Value:' label is followed by a slider ranging from 0 to 100, with the value set to 99. A 'Run Analysis' button is at the bottom.

### Step 2:

Select method of sub-setting. You can use the boxplot on the main panel to help choose the method. In the TCGA BRCA ds, IQR shows relatively larger spread in comparison to VAR and MAD.

### Step 3:

Choose a percentile cut-off to select the top most variable number of features (genes in this case). You can also choose cut-off based on the inclusion of a gene. Here 99<sup>th</sup> percentile i.e. top 1% shows increased variability.

Click Run button to display results in main panel each time any parameters are updated. The total number of selected genes are displayed in the main panel.

Genome- Wide analysis workflow is divided into 4 subparts: (1) Select most variable features (2) Construct a Heatmap of most variable features, (3) Identify Number of clusters and Assess Cluster Stability and (4) Identify Core Samples. Based on the core samples, heatmap is updated. Each feature can be used individually with the own user defined data. A Genome wide dendrogram can be quite dense and sometimes not necessarily informative. NOJAH displays an Interactive Genome-Wide dendrogram instead based on different normalization and scaling methods. User can also choose between the eight-different distance and seven different agglomerative linkage methods (Description on page #3).

Most-Variable Features

Heatmap

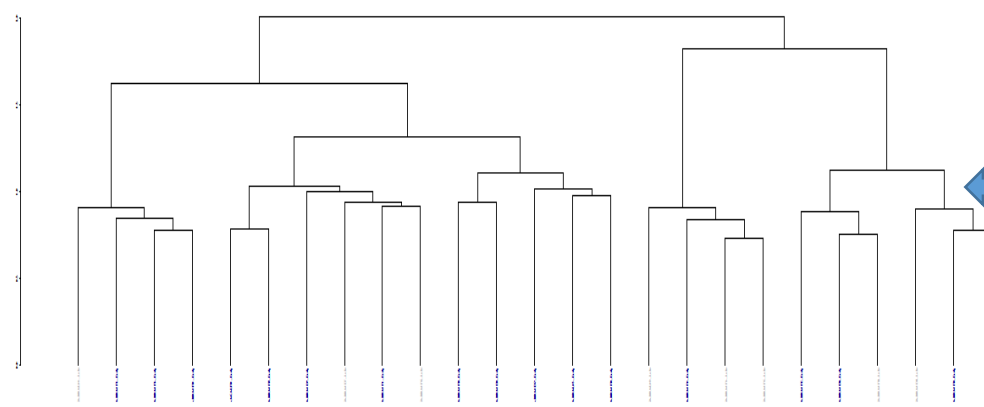
Cluster Number and Stability

Core Samples

Updated Heatmap

Workflow

Genome-Wide Dendrogram



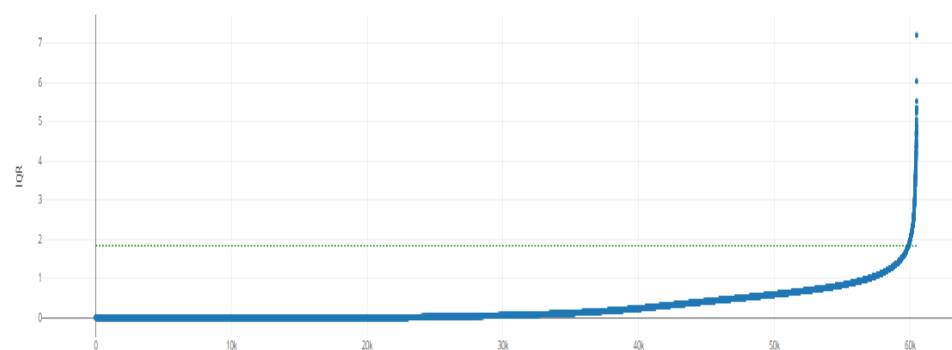
Measures of Spread



Ignoring the outliers, IQR shows the most spread based on the boxplots for the TCGA BRCA dataset.

## Select Most Variable Features

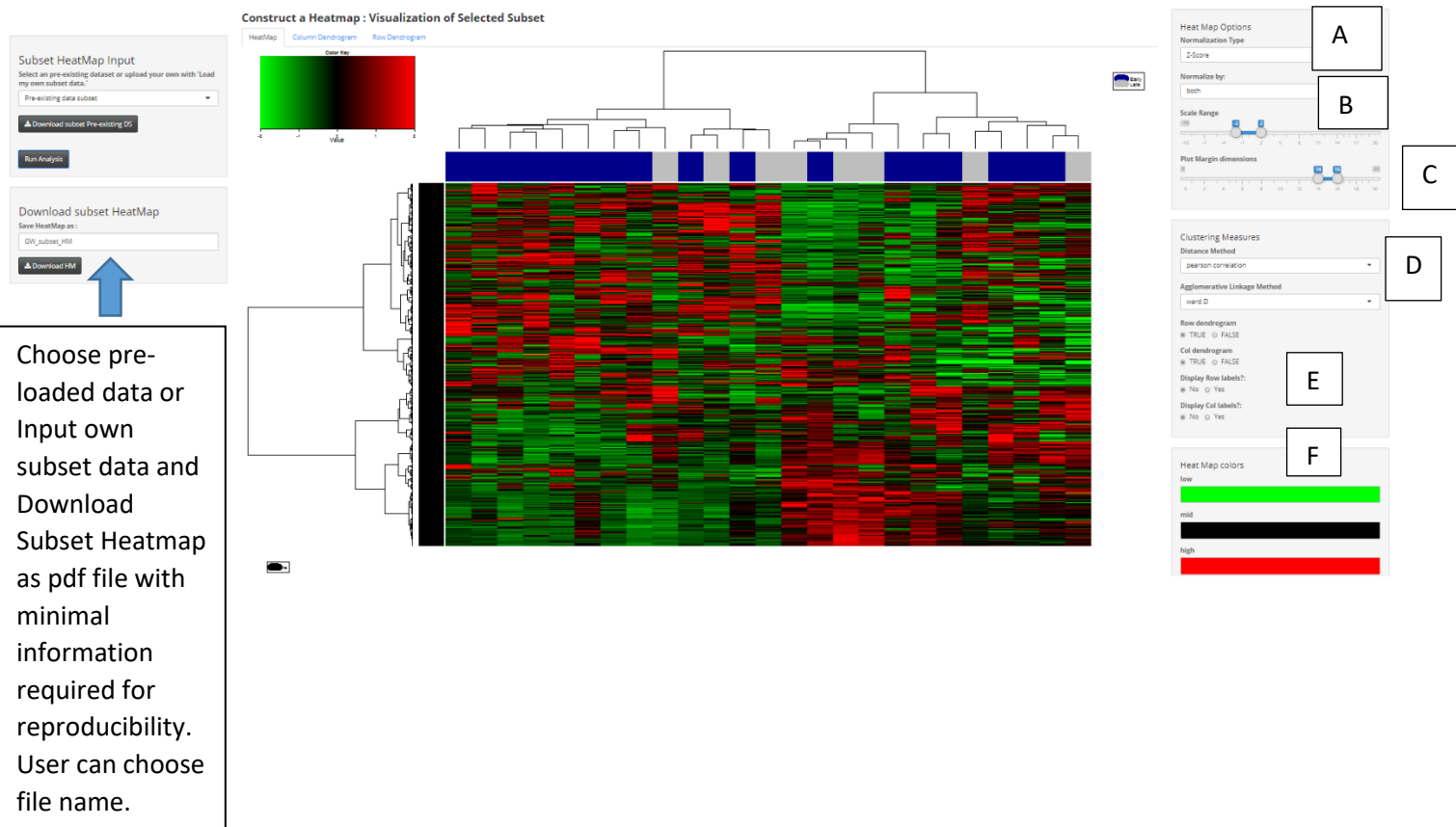
To see the position of your gene of interest, use the 'Choose Option' drop down to the right



The number of genes selected : 605

Most variable genes selected based on a percentile or row inclusion criteria. Specific gene can be chosen from drop-down and highlighted in orange on plot. The number of genes included are displayed.

Identify Most Variable Features



Interactive HeatMap of the top most variable genes selected using the above criteria. Separate tabs display column and row dendrograms. (See tutorial for part C for detailed run-through of heatmap options, Page #9)

- A. Data is z-scored before input into the heatmap.2 function.
- B. Data can be normalized by row, column or both, default: both
- C. Scale is set from -2 to 2 but can be changed by the user
- D. Choose clustering and distance measure, pearson and ward.D respectively are defaults
- E. Supervised row-wise or column wise clustering can be selected using FALSE option. Row and column labels can be displayed using TRUE option
- F. Change color of HeatMap by clicking on the high, mid and low colors

Details on heatmap options are available on page #12.

# Identify number of clusters and Access cluster stability

Most-Variable Features

Heatmap

Cluster Number and Stability

Core Samples

Updated Heatmap

Workflow

Consensus Clustering Input

Select an pre-existing dataset or upload your own with "Load my own subset data."

Pre-existing data subset

Download subset Pre-existing DS

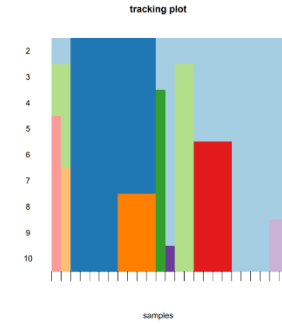
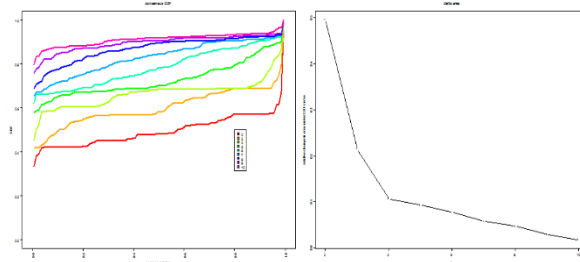
Run Analysis

Download Consensus Cluster Results

Download Consensus Cluster Results

Download results are available. Downloading the results will be automatically open.

Number of clusters and Assessing cluster stability



Consensus Clustering Options

Distance Measure: pearson

Clustering Method: average

No. of iterations: 10, 100, 500, 1000

Choose Optimal Number of clusters

Optimal k is: 2

Choose pre-loaded data or Input own subset data. Data format should be same as that in the preloaded example data (available for download) and Download Consensus cluster output for consensus heatmap as available from 'ConsensusCluster' Plus Bioconductor package.

Consensus CDF, Delta area plot are from the output results of the 'ConsensusClusterPlus' package. Along with the consensus matrix heatmap (available for download), they will help the user determine the optimal number of clusters in the data.

For this TCGA BRCA Expression data, two optimal clusters are predicted. The user can change the optimal clusters using the right panel.

Choose distance and clustering measures to perform **consensus clustering** using the 'ConsensusClusterPlus' package, to predict optimal number Sample clusters for the data.

Static parameter settings used:

- item resampling = 80%
- gene resampling = 80 %
- maximum evaluated k = 9
- clustering algorithm = Agglomerative Hierarchical clustering algorithm "hc"
- Same clustering method is applied to both inner Linkage and final Linkage parameters.

Most-Variable Features

Heatmap

Cluster Number and Stability

Core Samples

Updated Heatmap

Workflow

The larger the silhouette width, the better the stability. Samples with negative silhouette width signify poor cluster stability and are removed to create the core sample set.

# Identify Core Samples

Silhouette Input

Select an pre-existing dataset or upload your own with "Load my own subset data."

Pre-existing data subset

Download subset Pre-existing silhouette DS

Select an pre-existing cluster or upload your own with "Load my own clusters."

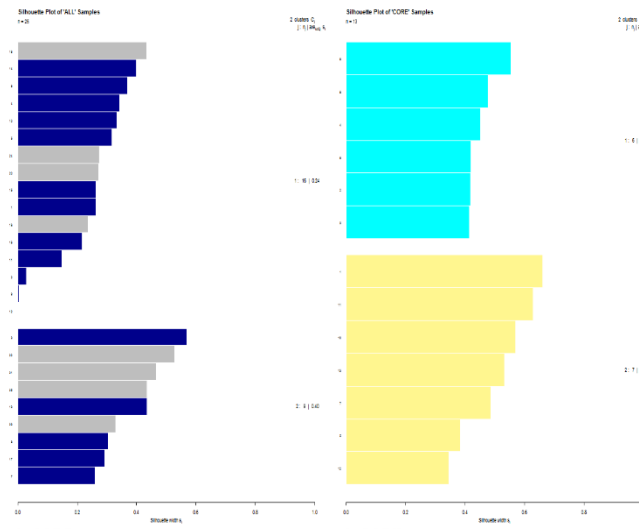
Pre-existing data subset

Download subset Pre-existing silhouette clusters

Run Analysis

Download Silhouette Core Samples

Download Silhouette Core Samples



Silhouette Options

Remove samples based on:

Fixed value

Change point

Remove Samples within Cluster1 with Silhouette width less than:

Remove Samples within Cluster2 with Silhouette width less than:

Choose pre-loaded data or own subset data and cluster classification. Data and cluster classification data format should be same as that in the preloaded example data (available for download).

Silhouette Options include choosing up to which width samples should be removed within each cluster. For TCGA BRCA data, 0.3 is used as cutoff.

For user uploaded data, additional options to choose distance will be made available.

## Input Core Data

Select an example dataset or upload your own with 'Load my own Core data.'

Pre-existing data subset:

Run Analysis

## Core Sample based Downloads

Type the file name you would like to save subset data as:

Core\_subset\_data

Download Subset data

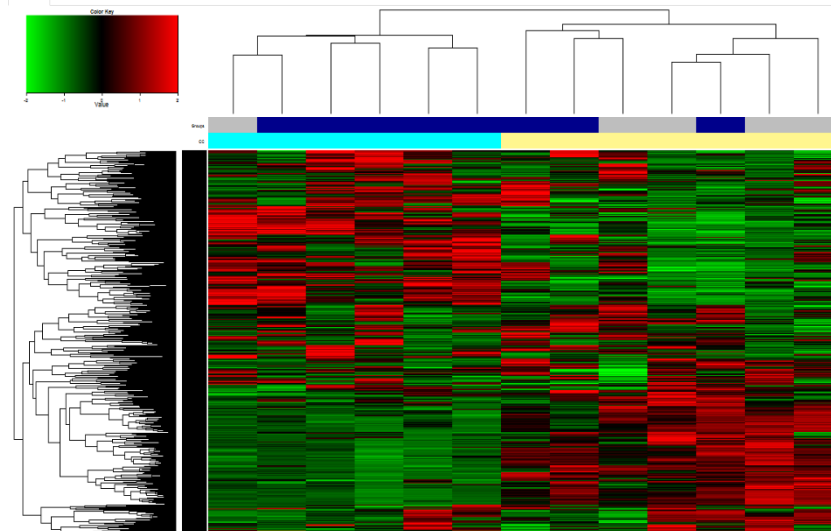
Save HeatMap as:

Core\_subset\_HM

Download HM

## Updated HeatMap based on Consensus Core Samples

HeatMap Column Dendrogram Row Dendrogram



## Heat Map Options

Normalization Type

2-Core

Normalise by:

both

Scale Range

0 100

Plot Margin dimensions

0 100

## Clustering Measures

Distance Method

euclidean

Agglomerative Linkage Method

complete

Row dendrogram

☒ TRUE ☐ FALSE

Col dendrogram

☒ TRUE ☐ FALSE

Display Row labels?

☒ No ☐ Yes

Display Col labels?

☒ No ☐ Yes

## Heat Map colors

low

mid

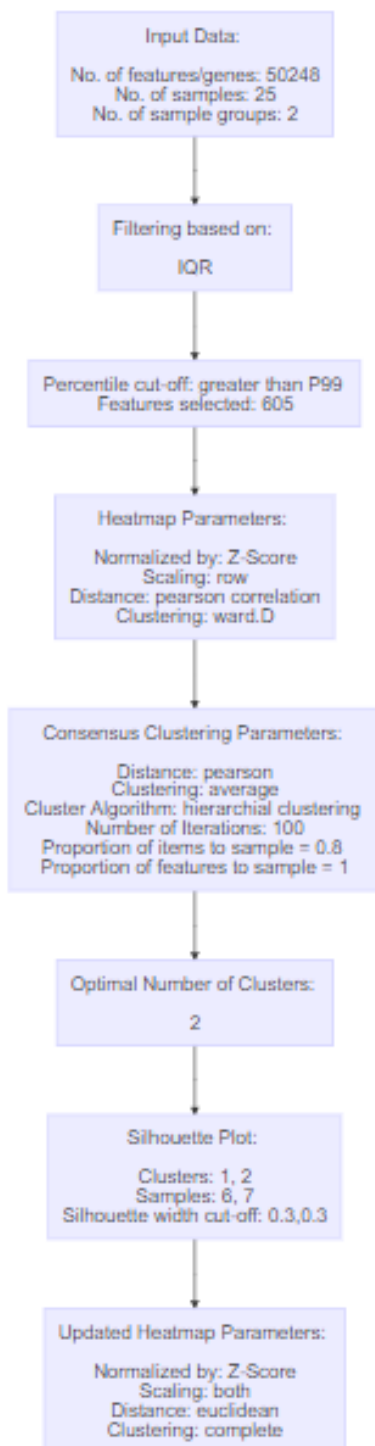
high

Download updated 'Core Samples' based Heatmap as pdf file with minimal information required for reproducibility. User can choose file name.

Interactive Heatmap of the top most variable genes for the Core Samples. Heatmap clustering is based on the consensus clusters (CC). The original sample clustering is available as the column color bar over the CC. Separate tabs display column and row dendrograms.

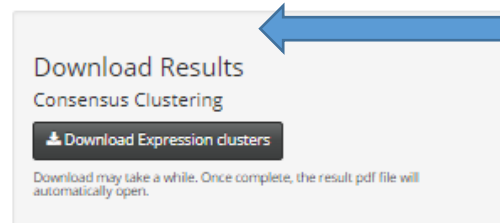
Options for heatmap are similar as shown on Page #3.

## Workflow for Genome-Wide Analysis



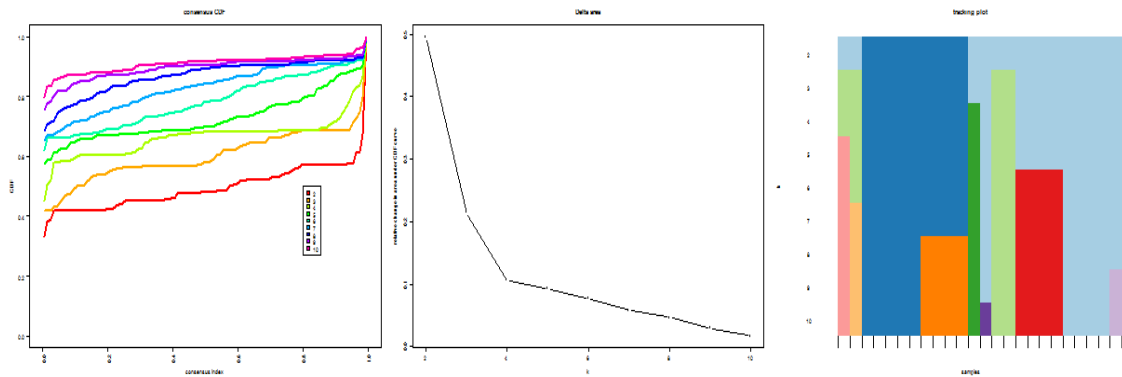
A complete workflow for the Genome-wide analysis is available in the workflow tab based on parameters selected on each tab. The workflow would be displayed for each tab in which the analysis was performed.

*\*Note: The same steps apply to Expression, Variant and Copy Number tabs.*



- item resampling = 80%
- gene resampling = 80 %
- maximum evaluated k = 9
- clustering algorithm =  
Agglomerative Hierarchical  
clustering algorithm “hc”
- Same clustering method is applied  
to both inner Linkage and final  
Linkage parameters.

## Determination of number of clusters by Consensus Clustering

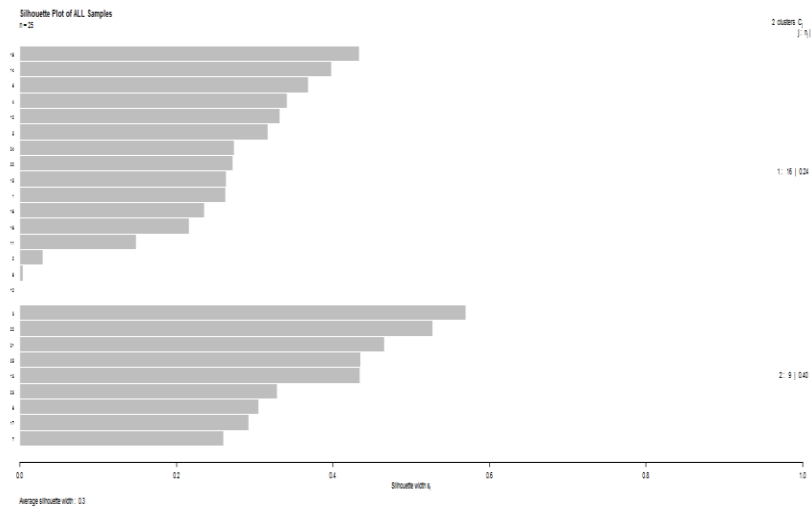


Choose Optimal Number of clusters

Optimal k is

2

## Silhouette Plot



These plots will be displayed using the parameter setting above like the GWH tab.

Consensus CDF, Delta area plot are from the output results of the 'ConsensusClusterPlus' package. Along with the consensus matrix heatmap (available for download), they will help the user determine the optimal number of clusters in the data.

For this Expression data, two optimal clusters are predicted. The user can change the optimal clusters using the right panel.

Silhouette Plot can further help confirm the identification of the number of clusters visually. The larger the average silhouette width, the more reliable the cluster structures are.

## Data type

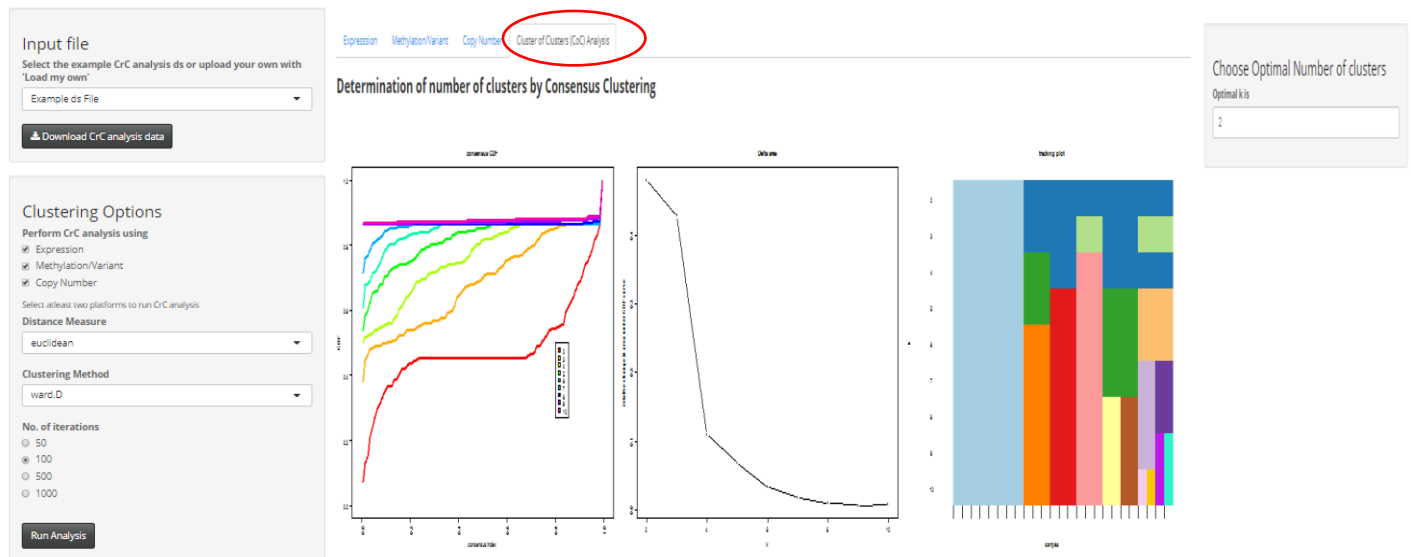
Choose data used

- ☒ Methylation  
☐ Variant

In the Methylation or Variant tab, the user can further select the data type used. This choice will be displayed in the CrC heatmap row label.



## CrC ANALYSIS



**Input file**  
Select the example CoC analysis ds or upload your own with 'Load my own'

Example ds File

Download CoC analysis data

**Clustering Options**  
Perform CoC analysis using

☒ Expression  
☒ Variant  
☒ Copy Number

Select atleast two platforms to run CoC analysis

Distance Measure  
euclidean

Clustering Method  
average

No. of iterations  
☒ 50  
☐ 100  
☐ 500  
☐ 1000

Run Analysis

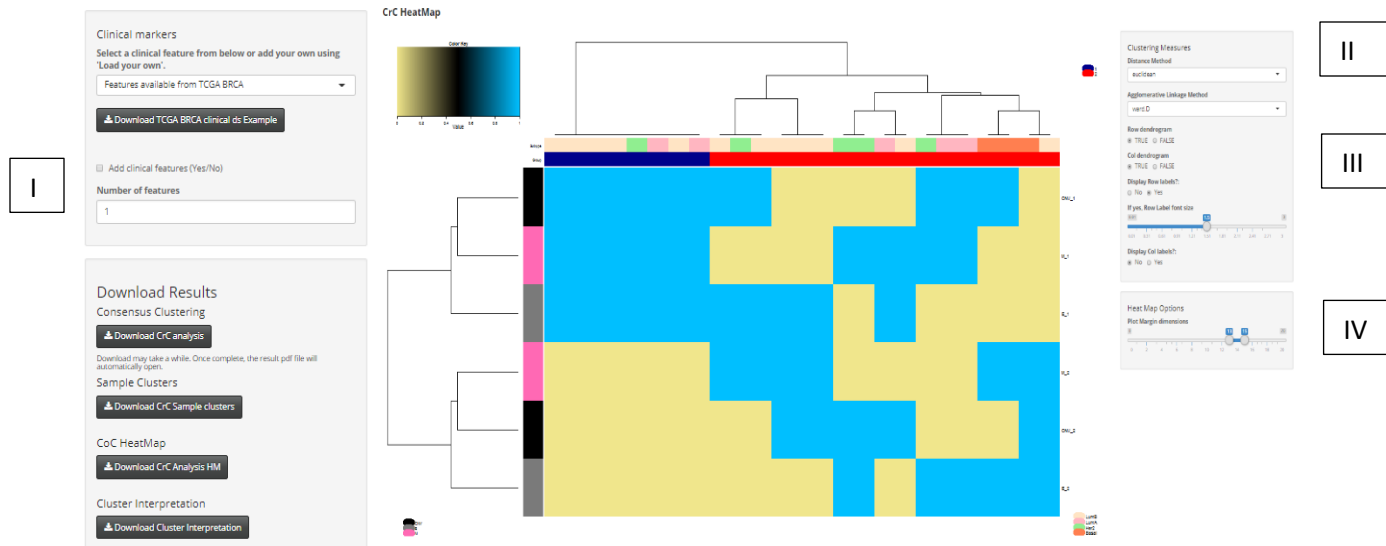
### Step 1:

Select the computed number of optimal RNASeq Expression, Methylation and CNV data from the previous 3 tabs or upload your own clusters using the load my own option. The user-uploaded cluster data should be in the same format as the example data. Example data is available for download. In addition, the same patients should be input in the same order for each platform.

### Step 2:

Select the platforms to base CrC analysis. User should select at least two platforms.

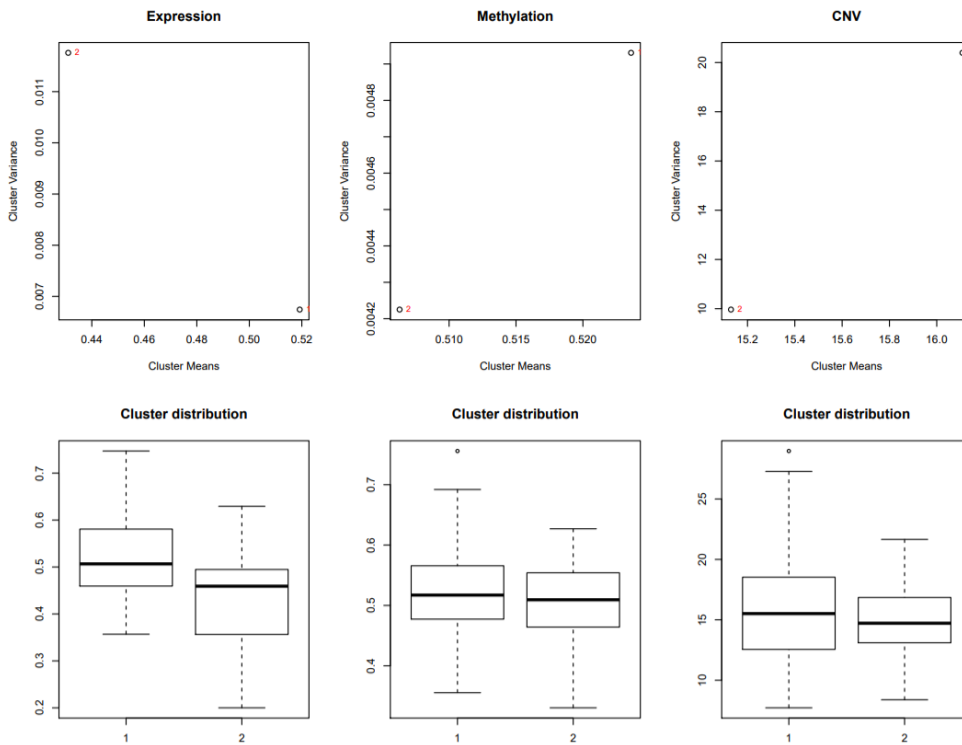
Select other parameter setting in the similar fashion to the previous tabs to run 'ConsensusClusterPlus' package.



### *Interactive HeatMap for Cluster of Cluster Analysis.*

1-0 Transformed matrix data based on the individual platform clusters is used as input into the modified heatmap.2 function.

- I. Add Single or multiple clinical feature(s) as bars just below the dendrogram. As an example, sample risk status is displayed above the predicted consensus cluster bar and can be downloaded using the download button.
- II. Choose clustering and distance measure
- III. Supervised row-wise or column wise clustering can be selected using FALSE option. Display Row and column labels using TRUE option. Adjust size of the labels using the slider.
- IV. Adjust Plot margins using the slider.



### Interpretation of CoC Analysis Cluster HM based on the individual platform clusters.

This option is available only when the actual expression, methylation or variant data is used. These plots will not be available if load my own option is used in the Cluster of cluster analysis tab.

Variance vs the mean plot of the lower triangular distance matrix serves as a relative measure of each cluster relative of the others within the same platform.

Boxplot of the individual clusters also helps determine which cluster has a relatively higher or lower median Expression (or median methylation or median CNV segment mean). The spread among the clusters is also informative.

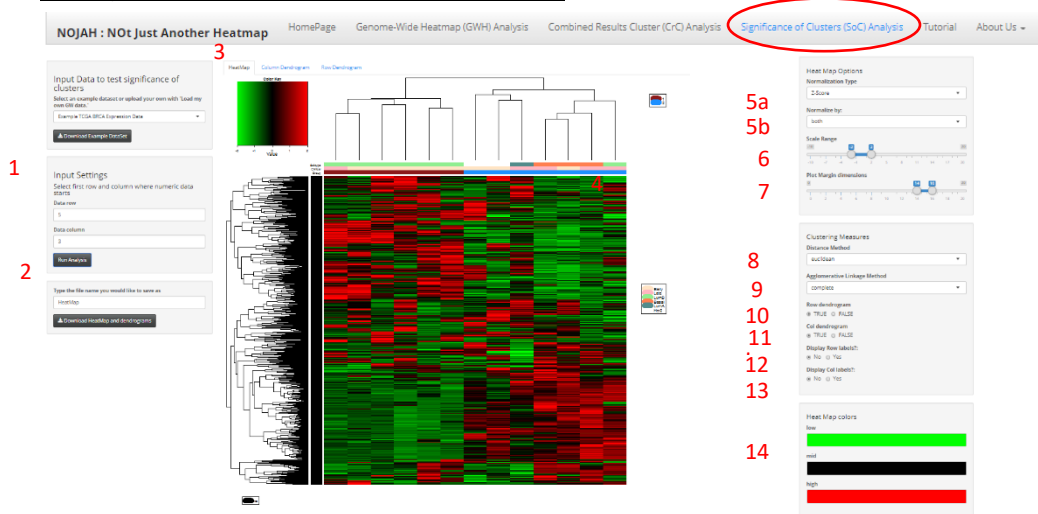
## Contingency Table(s)

Stratified by CNV

	CNV1M1	CNV1M2	CNV2M1	CNV2M2
E1	8	3	2	3
E2	3	2	2	2

Contingency table displays the distribution of samples among the clusters. When all three platforms are used, the contingency table is stratified by the third platform i.e. CNV.

## TAB C) SIGNIFICANCE OF CLUSTER ANALYSIS



**1:** Select dataset of interest. Using the dropdown, you can choose the example most variable TCGA BRCA (from GWH tab)/CoMMpass Expression dataset or upload your own. If uploading your own, format data in same format as in the example file. Also input numeric data start: row and column. In example data, numeric data starts on row 5 and column 3. Depending on each dataset, this needs to be adjusted.

**2:** Download example data using download button to view contents/formatting of example file.

**3:** If example file is chosen, Heatmap automatically displayed in the HeatMap tab.

**4:** HeatMap created using Z-score 'both row and column' normalization, 'Euclidean' distance and 'complete' agglomerative linkage method (i.e. default settings). Depending on dataset may take several minutes to load.

**5a, b:** Select a different normalization method you'd like for the data using drop down options. After choosing a different type, hit 'Run Analysis' button to update the heatmap.

The screenshot shows the 'Normalization Type' dropdown menu. The selected option is 'Z-Score'. Below the dropdown, there is a 'Normalize by:' section with a dropdown menu showing 'row|'. The dropdown menu is open, showing options: 'row', 'col', and 'both'.

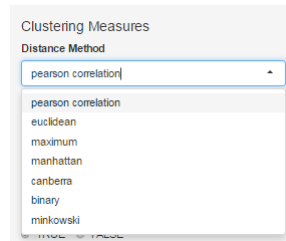
**6 (optional):** Drag slider to change scale range for the colors. Hit 'Run analysis' button to update Heatmap.



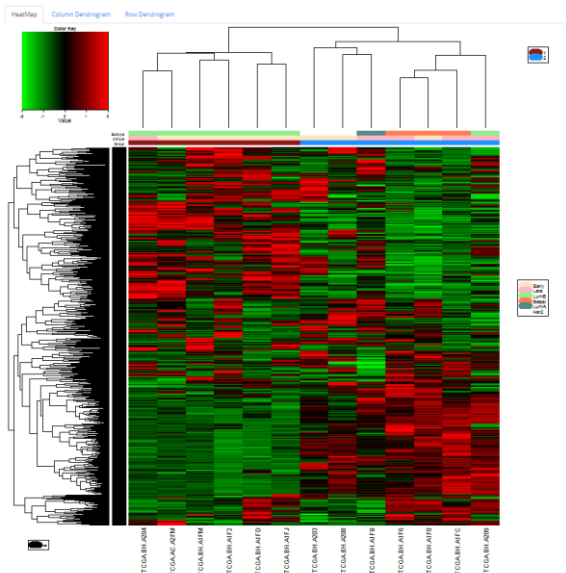
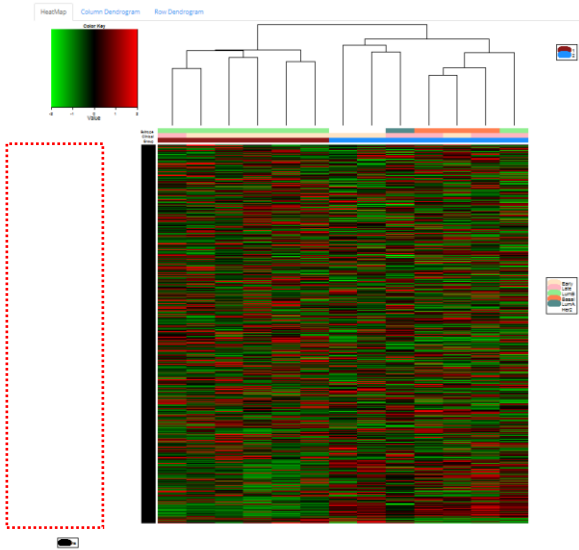
**7:** Select the Plot margins. If column dendrogram overlaps the legend, increase both margin points and vice versa until desired.



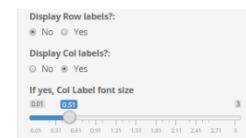
**8, 9:** Select Distance method and linkage method of choice using the drop-down options. Each selection will display modified heatmap.

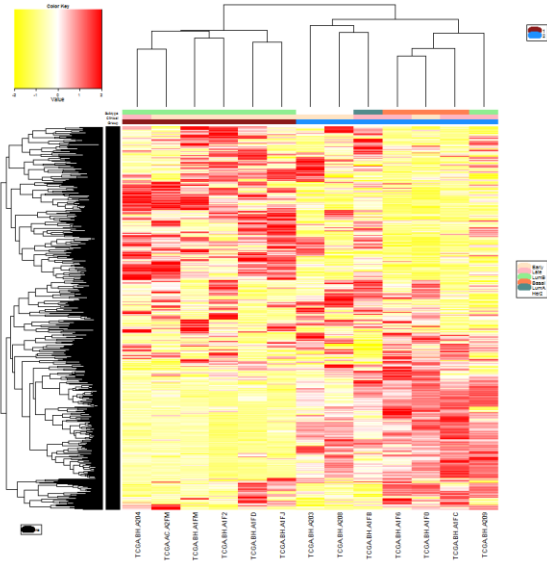


**10, 11:** Select either to display Row dendrogram or not. If FALSE is chosen, row dendrogram will disappear and data will not be ordered based on means. Same applies to Column dendrogram.



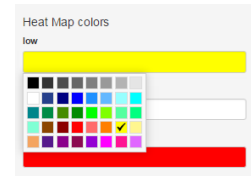
**12, 13:** Select Display Row labels = 'Yes' to see the corresponding genes. Additional slider appears to select, font size. Same applies to Sample labels.





**14:** Select color scheme. Red-Black-Green is typically used for Expression data and Blue-White-Red is used to represent methylation data. Heatmap will update as soon as color is chosen. After choosing desired color(s), click anywhere on screen to come out of color selection panel.

**15:** Input file name and click on Download button to save heatmap and the corresponding row and column dendrograms in pdf format as shown below using Chrome browser.



**16:** View in column dendrogram tab

**17:** Slider to adjust font size of the column dendrogram labels

**18 a, b:** a. Option to cut the tree. b. If yes is chosen, user is asked at which position they want to cut the tree (default at 2)

When selected, a table will appear that classifies Samples, their Groups, and their corresponding clusters.

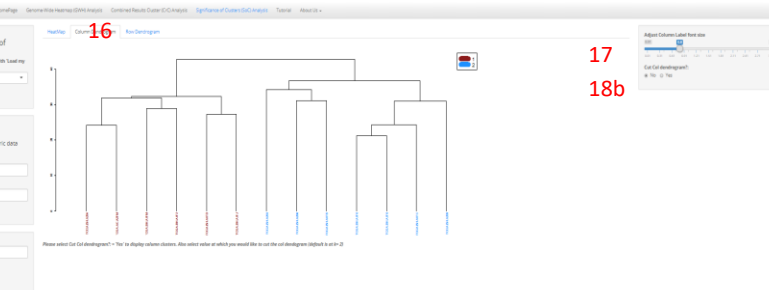
Use the drop down on upper left to display 5/10/All rows of the table.

**19:** Option to assess gene set significance in separation of the two clusters (Group1 vs Group2). Applicable only when >=2 clusters are available for analysis.

Assess Gene set significance in separation of specimens into 2 clusters?:  
☐ No ☐ Yes

When 'Yes' is selected, parameters for Monte Carlo p-value estimation will be made available.

20



**18a** Please select Cut Col dendrogram?: = "Yes" to display column clusters. Also select value at which you would like to cut the col dendrogram (default is at k=2)

Sample	Group	Cluster
1	TCGA-BH-A2A	1
2	TCGA-BH-A2B	1
3	TCGA-BH-A2C	1
4	TCGA-BH-A2D	1
5	TCGA-BH-A2E	1

**19** Would you want to assess gene set significance in the separation of specimens into two clusters? (Yes/No)

**19a**

**19b**

**19c**

**19d**

Select a dataset or upload your own with "Load my own data."

Example TCGA BRCA Exp Sampling Data

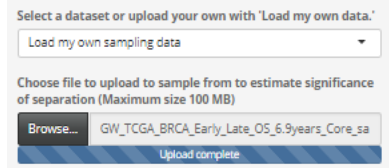
Sample size for bootstrap: 1000

No. of iterations for bootstrap: 1000

Go

Click the button to start sampling using bootstrap method for estimating the p-value. A progress indicator will appear shortly (~approx 10 seconds) on top of page indicating the status. Once complete, the p-value will be displayed in the main panel.

**19a:** Select Sampling dataset for bootstrap. An example GW TCGA BRCA Sampling data is available or user can input their own (up to 75 MB is allowed). Large CSV and TXT files can be converted to RDS file contain file size within 75 MB limit.

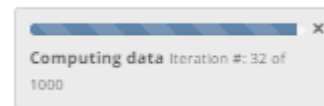


**19b:** Choose Sample size of the data for bootstrap. Use a size that does not exceed the original sampling data itself. For example, 1000.

**19c:** Select number of iterations you wish to perform. A good practice is to perform at least 1000 iterations for accuracy of analysis.

**19d:** Once all options are selected, press 'Go' button to start analysis.

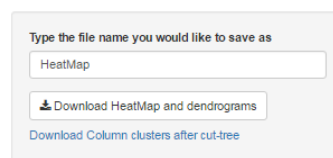
After approximately 10 seconds, a progress indicator will appear to track the time remaining for the analysis to be completed.



**p-value results** from the boot strap approach for calculation significance of clusters using Fisher's exact test will be displayed under the table along with the interpretation.

**20:** To download the p-value results as well, input the file names and click on Download button. The heatmap and the corresponding row and column dendrograms followed by the p-value results will be downloaded in pdf format.

To download the table for the classification of samples by clusters, click on link and the table will be saved as a CSV file.



Similar analysis can be performed on Row Dendrogram, provided you have at-least two row groups.